Part I: Machine Learning Methods on Real Data

1.1. Introduction

While a lot of conventional research has been conducted in the global setting that compares economic growth across countries, we hope to see if machine learning models can exploit empirical research findings to predict regional economic growth in the medium term. Recognising that social capital has a determinant effect on economic performance, we hypothesise that machine learning models could utilise social development indicators to predict economic growth (Iyer et al, 2005). Namely, we like to determine whether machine-learning methods could employ social development indicators as predictors for regional economic growth in the US. In this research, we assumed the inelasticity of social capital that a cross-sectional observation of social development is representative of the study period.

1.2. Data and Methodology

We collected data from the US Census Bureau for the economic and social indicators on the Census Tract Level made available from the American Community Survey 5-Year Data (ACS5). To avoid the complication of considering the effect of COVID-19, we used the 2014 and 2019 editions of the data. We retrieved the data by running API requests through an R session. Due to the computational demand, we restricted our research to the state of New York, which contains 4918 unique data points.

To measure the economic change, we retrieved the median income for both 2014 and 2019 and computed the percentage change over the time period. In addition, we created a new variable to classify a census tract based on the direction of change in median income. With this approach, we yield a dataset with a significant class imbalance where less than 20 percent of the census tracts recorded a decrease in median income. For the predictors, we identified opportunistically and obtained socio-economic indicators as measures of social capital, including education attainment and unemployment rate (*figure I in Appendix*). For observations with an “insufficient number of sample observations" for an estimated value, we imputed a value of 0 to preserve the dimension of the dataset. Filtering out data points with missing values, 134 census tracts were dropped. A preliminary exploration of the correlation heat map (*figure II in Appendix)* showed no strong correlation between any socioeconomic variables and a change in income.

The dataset was split into three sets: the training, testing, and validation set. We used the training data to fit the models for variable selection and hyperparameter tuning. The test data is used to compare and evaluate the models. Lastly, the validation set was used for the performance of the selected best model.

We first ran a simple multiple linear regression (MLR) model to help us understand the interaction between the predictors and the response variable. Assuming that the first model would have a high variance, we then proceeded to use the regularisation method using the L1 loss function, best subset selection and principal component analysis to compute three regression models. We also explore the use of neural nets for prediction. Based on the selected performance metrics, we identified the best training method and employed the best model to predict the change in median income on the validation set.

We then simplified the research question to a classification problem to see if the performance of machine learning methods would be better. We considered five approaches, Logistics regression, LDA, QDA, *k*-nearest neighbours classification and lastly tree-based method. We considered both the accuracy and the sensitivity-specificity trade-off to identify the best model and used it to predict the validation set.

1.3. Results

3.1 Regression Analysis

From the preliminary results on the training data where all predictors were used to train a MLR model, we observed some association between the predictors and the response variable. Most of the predictors however are insignificant. Nonetheless, it has a low R^2 (0.1198) even on the training set. We use this model to predict on the test data and it yields a RMSE of 0.2272.

3.1.1 Lasso Regression

With a cross-validation approach to tune for the best, $\lambda = 2.1488 \times 10^{-4}$, (*figure III in appendix*), we trained a model that uses 22 predictors. On the testing data, it got a RMSE of 0.2267.

3.1.2 Best Subset Model

Under the adjusted R^2 and C_p, they yield the models with similar complexity. On the contrary, *BIC* yields a model with lower complexity as it places a higher penalty on the number of predictors.*(figure IV in appendix*) We used the adjusted R^2 for our evaluation because we believed our model is fairly under-fitted given the low R^2 from the MLR. Under the R2 criterion, the best subset employs 16 variables and yields an RMSE of 0.2268.

3.1.3 Principal Component Regression

We selected the number of components in the PCR model by cross-validation(*figure V in appendix*) and one standard error rule. The resulting model is used to predict on the test data and it yields an RMSE of 0.1995.

3.1.4 Neural Network

We trained a neural network with two hidden layers on the training dataset(*figure VI in appendix*) and used it to predict on the test data. It took some time and the final model yields an RMSE of 0.2556617.

3.1.5 Comparison of the Regression Models

Regression Methods	Measure(Root MSE)	Time Taken(in seconds)
Linear Regression	0.2272314	0.0236630439758301
Lasso	0.2267227	0.0626819133758545
Best Subset	0.2268446	0.0244841575622559
Principal Component Regression	0.1995219	0.23395299911499
Neural Network	0.2556617	25.1943969726562

Figure 1: Evaluation Metrics of the Regression Model

Based on the RMSE, the PCR model performs the best. However, we recognised the significance of its computational demand. It would not scale well when studying all other US states. For the final model, hence, we selected the Lasso model which has the lowest RMSE and yields an RMSE of 0.2089 on the validation data.

3.2 Logistic Regression

Note that we used the best subset using accuracy as the criterion to choose the best model (instead of the sensitivity. This model chosen has 2 predictors. We train this model on the training data and use it to predict the test data. The model gives an accuracy of 0.83584. (misclassification error rate of 0.1642)

3.2.1 Linear Discriminant Analysis

We trained a LDA model and used this to predict on the test data. This model yields an accuracy of 0.8380355. (misclassification error rate of 0.1620)

3.2.2 Quadratic Discriminant Analysis (QDA)

We apply QDA on the training data and then use it to predict on the test data. We do observe the QDA is more sensitive to tracts with a decrease in economic performance. This model gives an accuracy of 0.7523511 (misclassification error rate of 0.2476).

3.2.3 K-Nearest Neighbour

Cross-validation method is used to select the optimal k for the KNN on the training data. Having selected this k, we use it to fit the test data. We do observe that this model takes some time for prediction and classifies all tracts to “up” label. The accuracy of this model is 0.8380 (misclassification error rate of 0.1620). This accuracy should be held in the context that there is a significant class imbalance and while the model performs well in terms of its specificity which is 1, we observe that the sensitivity is 0.

3.2.4 Random Forest

We first use a classification tree on the training data and observe that this tree always assigns an observation with an "up" label and fails to identify any census tract that has a decrease in median income (*figure VII in Appendix*). Since the "full tree" trained by the algorithm is close to a stump and the predicted outcome is always an increase in median income, we assumed that a simple tree is not suitable for this problem and did not proceed to prune the tree further. Instead, we move on to use a more complex tree-based model to see if the performance would improve.

We used Random Forest on the training data. Since the Random Forest is not sensitive to the number of trees we selected, we used the number of trees as 13 for computational efficiency. We used this model to predict on the test data and it yields an accuracy of 0.8391 (misclassification error rate of 0.1609)

3.2.5 Comparison of the Classification Models

Classification Method	Accuracy	Sensitivity	Specificity	Time Taken(in seconds)
Logistics Regression	0.8359457	0.0322581	0.9912718	3.29260611534119
Linear Discriminant Analysis	0.8380355	0.0516129	0.9900249	0.0218911170959473
Quadratic Discriminant Analysis	0.7523511	0.3806452	0.8241895	0.0124199390411377
K Nearest Neighbour	0.8380355	0	1	25.2629871368408
Random Forest	0.8390805	0.083871	0.9850374	2.86535620689392

Figure 2: Evaluation Metrics of the Classification Model

Although the simpler model yields a higher accuracy rate, they are insensitive to the census tracts with a decrease in median income. Given the class imbalance, we believe that the trade-off between sensitivity and specificity achieved in the QDA is superior to other models we explored, so we proceeded with the QDA as the final model. On the validation dataset, both the accuracy and sensitivity

dropped to 0.7346 and 0.2849 respectively. Nonetheless, it is still better than the performance of other models on the testing set in terms of sensitivity.

1.4. Discussion and Conclusion

The machine learning models applied to the dataset did not perform well on the data. The final regression model has an RMSE that is more than double the variance of the response variable. Similarly, the classification also performed poorly, especially in classifying census tracts that recorded a decrease in median income. This could simply be due to the noisy nature of socioeconomic data or due to the imbalance of classes which affects classification models like KNN. The performance of machine learning models on the dataset could also be restricted by the methodology.

Firstly, we assumed a linear relationship between all variables and the response variable. There could be a diminishing economic return on any given social capital and conversely an exponential relationship between the predictors and the response variable. There may also be an interaction between the socioeconomic variables (Iyer et al, 2005). As we observed the full linear regression model did not perform much worse than the shrunk models. This could indicate that the initial model has been misspecified and has a relatively low bias to start with.

Secondly, we only used 20 metrics to represent the social capital. Given the depth of the dataset and the potential to deepen the dataset by including more states, we could include more variables from the ACS5, which holds more than 4000 variables.

Thirdly, we assumed that social capital is static over the period of study. While this is a convenient assumption that facilitates the interpretation of the result, this social capital could also have changed in the time period, for example, due to migration which in turn affected the economic performance. A dominant example would be the downtown area of San Francisco. We believe that in future studies, an analysis of the changes in social capital over the time period would improve the rigour of the findings.

Fourthly, we did not adjust the median income by inflation using the same base year. We simply exploited the data available from a single credible source for our data. Adjusting for inflation with the same base year, it is likely to offset the class imbalance and a more ideal set-up for the classification problem.